*Ulster Med J* 2010;**79(2):**52-56

*Review*

# Metrics in Medical Education

Paul McCoubrie

Accepted 25 February 2010


## ABSTRACT

If every doctor is a teacher, then every doctor should be an examiner too. Assessment has a huge impact on learning; more so than most realise. Whilst there have been seemingly endless changes to current assessment strategies, there are some fundamental tenets to fair assessment that have changed little in recent decades. Similarly, whilst the hurdles to good quality assessment seem innumerable, there are lessons to be learnt from the literature that can lessen the impact of assessment on busy doctors.


## INTRODUCTION

"It is impossible to overestimate the importance of assessment"

David Newble, 1998

The word physician derives from the archaic noun *physic,* meaning the art or science of treatment with drugs or medication, whereas the word *doctor* originates from the Latin word (genitive case *doctoris)* for teacher. Indeed, countless generations of doctors have recognised the obligation to train others and have, more or less, happily done so since the inception of our trade a few millennia ago. More recently the General Medical Council (GMC) have formally reasserted the educational obligations of all doctors.[1]

I contend that all doctors should also be examiners. At first sight this statement may seem deliberately inflammatory; yet another unwelcome demand on busy medical practitioners. However I will explain that this is neither controversial nor onerous.

Of the twelve widely agreed roles of a medical teacher [2], the one that many doctors gloss over (or frankly ignore) is being an examiner. This is ironic as all doctors already formally and informally assess others; perhaps they don't recognise it as such. Such disparate tasks as interviewing for a new member of clerical staff, giving feedback to a trainee, planning a teaching session or formally examining medical students all entail the same principles of assessment.

This article, therefore, has three aspects. First, it will emphasise the importance of assessment. Second, it will examine obstacles to good assessment. Third, it will review the key issues in modern assessment, carefully distilled from the ever-expanding evidence base. The overall goal is to assist the reader to become more effective at assessment and perhaps to be realistic about what can and cannot be achieved.

## THE EDUCATIONAL IMPACT OF ASSESSMENT

"Teaching without testing is like cooking without tasting or writing without reading"

Ian Lang, 1991

Some doctors see exams as a necessary but time-consuming evil, a distraction from teaching and learning. However, in reality, assessment is not only intrinsic to any education endeavour but it is one of the most important tasks. This is simply because of the powerful effect of any assessment on the learner. If assessment is ignored or paid mere lip service then the teacher immediately lessens the impact of their teaching. Bizarre although it may seem, not assessing the learner does them a disservice.

Most assessment is relatively informal and low key. It is to check that learning has occurred, to reinforce particular important points and provide feedback to the learner to help them improve. This style of assessment is commonly known as *formative* assessment. This is in distinction to *summative* assessment which is typified by robust methods, lengthy tests and comparison to a pass / fail standard. Summative assessment includes formal examinations where decisions about career progression are made - so-called "high stakes" exams.

Many authors have documented the tremendous impact that high-stakes exams have on the learner [3]. Some authorities assert simply that "assessment drives learning" [4]. They state that students and trainees feel overloaded by work and hence they strategically learn what they perceive as necessary in the face of exams. From the student or trainees perspective, tests serve an additional, somewhat hidden purpose: they communicate what the "real" course goals and objectives are. Put metaphorically, "The assessment tail wags the curriculum dog", or, more crudely, "Grab students by the tests, and their hearts and minds will follow". [5]

Lambert Schuwirth of Maastricht University has coined the "law" of educational cause and effect. This states "for every evaluative action, there is an equal (or greater) (and sometimes opposite) educational reaction". For example, Newble & Jaeger showed in 1983 that if written testing was

University of Bristol, Department of Radiology, Southmead Hospital, Bristol BS10 5NB

Correspondence to: Dr Paul McCoubrie ,

paul.mccoubrie@nbt.nhs.uk

Tel 01173236341 Fax 01173235122

emphasised, then students focused on book-based learning, whereas if clinical testing was emphasised, students tended to focus on rehearsing their clinical skills on patients.[6]

There are several ways in which learning can be predictably affected. Assessment drives learning through its content, through its format, through the information given afterwards and through the frequency and timing of exams.[3] This effect on learning is often known as *consequential validity*.

The unpredictable side of Schuwirth's law arises because the relationship between assessment and learning is complex. Students and trainees learn subjects that are explicitly not examined [7]. What students actually learn is a very complex social phenomenon; a whole melange of tacit social, cultural and political issues that affect learning. Labelled the "Hidden Curriculum" it was first directly addressed by Benson Snyder in 1971.[8] It can represent a substantial portion of learning. In one study, 75% of final-year medical students sought extracurricular teaching.[9]

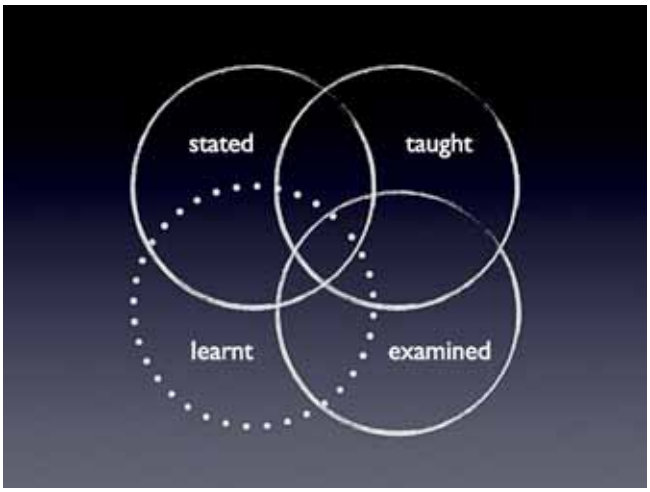


*Fig 1*. Conceptual map of the Hidden Curriculum

These models of learning can be illustrated as a Venn diagram of overlapping circles representing different ways of looking at a teaching program or curriculum (Figure 1). First, there is a formally *stated* curriculum, often written and widely available. This varies in style, content and format. Second, there is a *taught* curriculum; the subjects covered in teaching sessions. Third, the *examined* curriculum is that covered by assessment processes. Lastly, there is the *learnt* curriculum, the enigmatic and slightly unpredictable subjects that students and trainees actually learn. Of note in this model, the first three are under direct control of the teacher but the latter is not. One always hopes the *learnt* curriculum will overlap significantly with the others. A particularly well-organised teacher will have tight overlap between the stated, taught and examined curriculum, hence making it likely the learnt is too. But the examined curriculum is the one that is mostly likely to have overlap with the learnt curriculum. Perhaps the most important take-home message here is that assessment steers learning and the canny teacher harnesses assessment to do just that.

## TRADITIONAL FUNDAMENTALS OF GOOD ASSESSMENT

Assessment's primary role in high-stakes exams should be that of a gold standard test in the diagnosis of incompetence: a test that really sorts the wheat out from the chaff. However in formative tests, the focus is on informing personal development. This doesn't mean that formative assessment should be cursory or brief. Quite the opposite, good quality feedback needs good quality data.

Whilst practical constraints often limit assessment, as a principle it should be appropriate and proportionate. For example, if the purpose were to inform an individual that they have reached appropriate levels of expertise in a particular procedure, it would be inappropriate to set a gruelling written exam. However, such a rigorous and searching written exam would be a perfectly acceptable way of testing knowledge in a formal and important setting such as medical school finals.

Irrespective of its purpose, a good test should follow established methodology. Historically, the focus on a good test was adequate *metrics* within bounds of feasibility; that is mainly achieving a highly reliable and valid test but also one that is easily administered.

**Reliability** is a fairly straightforward idea: it is the degree to which a test consistently measures whatever it measures. It is a statistical concept, where a stated reliability coefficient or "r-value" is expressed where 0 is zero reliability and 1 is total reliability. Reliability improves with increasing the length of test, where the spread of scores is broad and even, where the level of difficulty is moderately high and the objectivity of marking is high.[10] Reliability can be calculated in a number of ways but the key message is that r=0.8 is an acceptable level for high-stakes exams.

**Validity** is a complicated concept in educational testing. Simply put, an exam is valid when it measures what it is supposed to measure. This is not a yes / no answer but a degree to which supporting evidence has been produced, or to what degree a theoretical premise supports an interpretation. The modern view is that validity is a single unitary construct with different aspects.[11] To be considered valid, an assessment should: -

- *Sample widely*. This is possibly the most important aspect. Doctors do not perform consistently from task to task. Hence a valid test samples systematically and representatively across what is supposed to be measuring. To do so usually makes a test long. When there is demonstrable evidence, this can be called content validity. Without evidence, it can be called face validity, a poorer measure
- *Differentiate*. It should be able to differentiate between groups of known differences. This can be called construct validity.
- *Agree with other tests*. If the results correlate well with another well-established test, it is said to have good concurrent validity.
- *Predict future performance*. Whilst most tests are administered to find something about future behaviour,

Table 1.

*Comparison of different methods of assessment (after Augustine et al [35]).*

| Assessment | Reliability | Validity | Feasibility | Acceptability | Educational effect |
|---|---|---|---|---|---|
| Multiple choice question | +++++ | + | +++++ | + | Makes trainees revise from written sources |
| Complex written (i.e. short notes) | ++++ | ++ | ++++ | ++ | Written sources are favoured but with less emphasis on facts |
| Oral exam | ++ | ++ | ++ | +++ | Trainees rehearse oral skills |
| Practical skill simulation | +++ | ++ | +++ | ++ | Encourages trainees to practice on models |
| OSCE or short case | ++ | +++ | ++ | +++ | Mixed effect; skills are rehearsed but can lack context |
| Long case | ++ | +++ | ++ | +++ | Trainees rehearse total performance |
| Workplace-based assessment | ++ | ++++ | ++ | ++++ | Focuses attention on clinical performance |
| Video assessment | ++ | +++++ | + | +++ | Trainees rehearse being recorded |
| *In-cognito* simulated patients | ++ | +++++ | + | +++++ | Revision emphasizes communication skills |

tests rarely perform well as hoped. Furthermore the longer that elapses, the poorer the correlation becomes. This can be called predictive validity.

- *Be real (or very realistic)*. Most testing methods aim to simulate reality but this is clearly second best to testing clinical performance in real life. Simulations and written testing should aim to mimic real clinical practice closely.
- *Guide learning*. As above, consequential validity is crucial.

Having said all this, it is virtually impossible to find a measure that is simultaneously fully valid, highly reliable yet feasible. When the inevitable compromises are made, then validity must remain the number one consideration. A comparison of the commonly used different methods of assessment is given in Table 1.

## CONTEMPORARY ISSUES

The focus on adequate metrics and feasibility has moved on a little in recent decades.

### *Fairness*

A fair or authentic exam is a defensible exam. Naturally it should be reliable and valid. In addition, questions should be carefully constructed by experienced examiners and reused with care. Adequate standard setting is also crucial. There are three main ways of setting a pass mark: *holistic, norm-referenced* and *criterion-referenced*.

A holistic model is simplicity itself, involving a fixed pass mark. Obviously the arbitrary nature of this is unreliable and is not recommended. In norm-referencing, the standard is based on the performance of the group being assessed. It is a relative pass mark and thus varies from group to group. Norm-referencing is quick and can be useful for formative assessment. Criterion-referencing refers to an absolute standard, irrespective of the group and is preferred for summative assessment.[12] It is worthy noting that criterion-referencing is relatively laborious. It has also several educational connotations regarding test construction.[13] Furthermore, many "criteria" are based on judgements of individuals or a small group, hence criterion-referencing is not without its critics.[14]

### *Workplace-based Assessment*

Despite improving fairness of traditional medical assessments, they have inherent deficiencies. The recurrent criticism centres around validity; results of traditional tests do not necessarily correlate with what doctors can actually do in their everyday practice.[14, 15] To allow more valid assessment, a number of assessment tools for use in the workplace have become available. These attempt to retain the authenticity of apprentice-style learning and assessment but adapted to modern working patterns. Instead of one master assessing a trainee, snapshots of the trainee in the workplace are taken to build up an accurate picture of their competence. Workplace-based tools enable the following:

- *Multiple perspectives.* A complete assessment of an individual can be achieved by gathering multiple perspectives of professional practice.[16] The use of multiple methods and assessors reduces bias and thus an accurate picture of the trainee is built up like a pointillist painting.[17]
- *Total practice assessment*. Many important aspects of professional practice such as abilities to work in a team,

 

teach, research and communicate are currently not seriously or formally assessed, mainly as they have been considered difficult to assess rigorously.[18, 19] These generic skills cannot be judged against a simple preordained standard nor can they be quantified easily, but must utilise qualitative and descriptive information. Such assessment is challenging as it relies on professional judgment of the assessor to make decisions regarding the trainee's performance without compromising objectivity.[20]

- *Charting of competence development*. The notion of suddenly becoming "qualified" to do something is illogical.[21] Competence is greyscale not black or white. It is acquired slowly rather than in one sudden epiphanic moment. Accumulation of data over time enables the development of competence to be charted. This can be conceptualised as assessment being a hurdle race rather than a high jump competition.

*The Modern Role of Written Testing*

The pre-eminence of written testing methods has been questioned. For example, one persistent criticism is that doctors do not answer batteries of complex MCQs in their day-to-day work, yet MCQs feature heavily in exams. The same argument runs that MCQs and other written question formats are therefore not particularly valid. However MCQs are the most time-efficient written test format, hence reliable testing is made feasible. MCQs also allow broad sampling of content that is unachievable in most other testing formats, particularly when dealing with large numbers of students or trainees. This achieves high content validity. Furthermore, they make learners hit the books, swotting up on book-based knowledge. If this is a desired activity, then they have good consequential validity.

One issue that is very clear from the literature is that the one single factor that predicts expertise is knowledge.[22] It follows that assessing knowledge using a written test is a perfectly reasonable way of assessing expertise. So, whilst MCQs lack acceptability and have some validity issues, they are good at testing knowledge, hence one's expertise. The way to improve their validity is to combine MCQs with a more practical or clinical exam to encourage broad learning.

## OBSTACLES TO ASSESSMENT

There are many potential reasons why many doctors feel uncomfortable assessing others.

*Lack of Training*

A lack of training in assessment is a common finding, both at an undergraduate[23] and a postgraduate level.[24] A survey of 529 hospital consultants found that 88% were involved in teaching but only 34% had any teacher training. The majority (67%) indicated that they needed training in assessment and appraisal skills.[25] Another survey of 441 hospital doctors found that, "giving feedback constructively" and "assessing the trainee" were two of the top three most commonly stated themes in which they would like more training.[26]

*Time & Resource Constraints*

These are ubiquitous in the era of ever-increasing NHS workloads. Further factors demand non-existent time and resources: a 50% increase in UK medical student numbers since 1996; a lack of senior trainees due to the legal constraints of working hours, together with all the challenges of teaching today's generation - "Generation Me".[27] Inevitably the motivation to improve assessment in the UK relies too heavily on the altruism of individuals. John Bligh notes that there are many "well-meaning, earnest teachers facing day-to-day practical problems in full awareness of what should be done, but only too aware of what can be achieved in the circumstances".[28]

*Tradition*

The current generation of medics have grown up on a steady diet of tests, often sitting up to 100 separate high-stakes examinations in their teenage and adult life. They are unsurprisingly test-weary, with a potential significant toll on their professional health.[29] Senior trainees and established medical practitioners are appropriately cynical about assessment but surprisingly accepting of unfair testing. Perhaps this is realism: whilst learners can walk away from bad teaching, assessment is usually mandatory irrespective of its quality. However, the same individuals are highly test-wise. This can be used to an advantage as their perceptions of an exam, its authenticity and overall fairness are valid and should be sought in any evaluation process.[30]

*Technical Complexity*

The number of scientific publications on assessment over the last decade has mushroomed. There has been an explosion in the number of proposed instruments, each with its unique TLA (three-letter acronym). The educational literature can be difficult to access and the technical jargon of psychometrics (the study of educational measurement) can further discourage casual browsing.[3] As a result, educational institutions are finding that they need staff with technical knowledge and understanding of assessment issues who can provide guidance.[2] The inaccessibility and complexity of these issues can prove daunting to even the most enthusiastic medical teacher.

*Appraisal*

Modern performance review tends to blur the boundaries between appraisal and assessment. The two are related but fundamentally different processes. Assessment is an explicit objective evaluation against defined criteria. Appraisal is a confidential, supportive review process of individual and institutional needs. Although appropriate assessment can inform appraisal processes, appraisal outcomes should not inform assessment.[31] Unfortunately, this is the precise basis upon which the GMC plans to base revalidation processes.

*Procrastination & Grievances*

Dealing with a student or trainee in difficulty can be so problematic that, "...it is far too easy to just pass the trainee and let someone else deal with the problem".[24] Freidenberg recognises this procrastination, leaving this "weeding out" to the certification board, possibly to avoid exposure of inadequate documentation at grievance hearings.[32] Where a student or trainee fails to progress satisfactorily, withdrawal from the programme can be recommended. However, the legal challenge to such dismissal can be extreme; Tulgan et al give

 

an example where an aggrieved resident mounted a 9-year legal test of dismissal policies, culminating in an appeal to the United States Supreme Court.[33] However, legal challenges to reliable and valid exams have generally been unsuccessful.[34]

## CONCLUSION

If every doctor is a teacher, then every doctor should be an examiner too. Assessment has a huge impact on learning; more so than most realise and it can be deliberately used to improve learning. Whilst there have been seemingly endless changes to assessment methods and strategies, there are some fundamental tenets to fair assessment that have changed little in recent decades. Similarly, whilst the hurdles to good quality assessment seem innumerable, there are lessons to be learnt from the literature that can lessen the impact of assessment on busy doctors.

The author has no conflict of interest

## REFERENCES


1. General Medical Council. The Doctor as teacher: archived policy document.. London: General Medical Council; 1999. Available online from: http://www.gmc-uk.org/education/postgraduate/doctor_as_teacher.asp. Last accessed April 2010.
2. Harden RM, Crosby J. AMEE Guide No 20: The good teacher is more than a lecturer - the twelve roles of the teacher. *Med Teach.* 2000; **22(4)**: 334-47.
3. van der Vleuten CP. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ.* 1996; **1(1)**: 41-67.
4 Wass V, van der Vleuten C, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001; **357(9260)**: 945-49.
5. Swanson DB, Case SM. Assessment in basic science instruction: directions for practice and research. *Adv Health Sci Educ.* 1997; **2(1)**: 71-84.
6. Newble DI , Jaeger K. The effect of assessments and examinations on the learning of medical students. *Med Educ.*1983; **17(3)**: 165–71.
7 McLachlan JC. The relationship between assessment and learning. *Med Educ.* 2006; **40(8)**: 716-7.
8. Synder BR. The hidden curriculum. New York: Knopf; 1973.
9. Goodfellow PB, Milton RS. Extra-curricular tutoring. *Med Educ.* 2001; **35(10)**: 1001.
10. Newble DI. Assessment. In: Jolly BC, Rees L, editors. Medical education in the millennium. Oxford: Oxford University Press; 1999. p. 132-42.
11. Messick S. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am Psychol.* 1995; **50(9)**: 741-9.
12. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences.. 3rd ed. Philadelphia: National Board of Medical Examiners; 2001.
13. Ricketts C. A plea for the proper use of criterion-referenced tests in medical assessment. *Med Educ.* 2009; **43(12)**: 1141-6.
14. Levy A. Personal view: The educationalists' standard stranglehold. *BMJ.* 2009; **338(7692)**: b690.
15. Miller GE. The assessment of clinical skills/competence/ performance. *Acad Med* 1990; **65(9 Suppl)**: S63-7.
16. Rethans JJ, Norcini JJ, Baron-Maldonado M, Blackmore D, Jolly BC, LaDuca T, Lew S, *et al.* The relationship between competence and performance: implications for assessing practice performance. *Med Educ.* 2002; **36(10)**: 901-9.
17. Schuwirth LWT, Southgate L, Page GG, Page GG, Paget NS, Lescop JM, Lew SR, *et al*. When enough is enough: a conceptual basis for fair and defensible practice performance assessment. *Med Educ.* 2002; **36**: 925-930.
18. Davies H, Howells R. How to assess your specialist registrar. *Arch Dis Child.* 2004; **89(12)**: 1089-93.
19. Epstein RM. Assessment in medical education. *N Engl J Med.* 2007; **356(4)**: 387-96.
20. van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ.* 2005; **39(3)**: 309-17.
21. Fish D, Coles C. Medical education: developing a curriculum for practice. Maidenhead, UK: Open University Press; 2005
22. Glaser R. Education and thinking: the role of knowledge. *Am Psychol.* 1984; **39(2)**: 93-104.
23. Fowell SL, Maudsley G, Maguire P, Leinster SJ, Bligh, J. Student assessment in undergraduate medical education in the United Kingdom 1998. *Med Educ.* 2000; **34(Suppl. 1):** 1-49.
24. Long G. Documentation of In-training assessment for radiology trainees. *Clin Radiol.* 2001; **56(4)**: 310-5.
25. Gibson DR, Campbell RM. Promoting effective teaching and learning: hospital consultants identify their needs. *Med Educ.* 2000; **34(2)**: 126-30.
26. Wall D, McAleer S. Teaching the consultant teachers: identifying the core content. *Med Educ*. 2000; **34(2)**: 131-8.
27. Twenge JM. Generational changes and their impact in the classroom: teaching Generation Me. *Med Educ.* 2009; **43(5)**: 398-405.
28. Bligh J. Assessment: the gap between theory and practice. *Med Educ.* 2001; **35(4)**: 312.
29. Gunderman RB. The perils of testing. *Acad Radiol.* 2001; **8(12)**: 1257-9.
30. McCoubrie, P. Improving the fairness of multiple-choice questions: a literature review. *Med Teach.* 2004; **26(8)**: 709-2.
31. Archer J. Assessments and appraisal. **In:** Cooper N, Forrest K, editors. Essential guide to educational supervision in postgraduate medical education. Chichester: Wiley-Blackwell; 2009. p 107-22.
32. Freidenberg RM. An endangered art: teaching. *Radiology*. 2000; **214(2)**: 317-9.
33. Tulgan H, Cohen SN, Kinne KM. How a teaching hospital implemented its termination policies for disruptive residents. *Acad Med.* 2001; **76(11)**: 1107-12.
34. Tweed M, Miola J. Legal vulnerability of assessment tools. *Med Teach.* 2001; **23(3)**: 312-4.
35. Augustine K, McCoubrie P, Wilkinson JR and McKnight L. Workplace-based assessment in radiology - where to now? *Clin Radiol.* 2010; **65(4)**: 325-32.


 